第 39 卷第 9 期
录用定稿
计算机应用研究
Application Research of Computers
Vol. 39 No. 9
Accepted Paper

基于离散泊松混合模型的教学评价数据建模 *

黄　浩，颜　钱，甘　庭†，李石君
(武汉大学 计算机学院，武汉 430072)

摘　要：分析学生在教学评价系统中对于教师的评价数据有助于教师了解学生对授课教师的真实态度，总结教学经验，改进后续的教学方式，提高教学质量。但是进行教学评价时，学生中可能会出现随意评价或者恶意评价等问题，导致评价数据中包含大量噪声，造成反馈数据的不理想。因此，提出了一种离散泊松混合模型来对包含噪声的学生的评价数据进行建模，将混合模型中的每一个离散泊松分量对应一类具有相似评价模式的学生，借由离散泊松分布中的模型参数来表示对应评价模式中的评价分数。通过构建对数似然函数来衡量混合模型和评价数据的拟合程度，采用梯度下降的方法求解拟合程度最高的模型参数，找到学生对于教师的真实评价，保证教学评价系统中师生间的有效沟通。大量实验结果表明模型能够快速准确地从含有噪声的评价数据中识别出具有不同评价模式的学生，掌握学生对于教师的真实评价情况。

关键词：教学评价系统；众包思想；泊松混合模型；参数估计方法
中图分类号：TP311　　doi: 10.19734/j.issn.1001-3695.2022.01.0042

Teaching evaluation data modeling based on discrete Poisson mixture model

Huang Hao, Yan Qian, Gan Ting†, Li Shijun
(*School of Computer Science*, *Wuhan University*, *Wuhan* 430072, *China*)

Abstract: Analyzing the evaluation data of students to teachers in the teaching evaluation system helps teachers understand the true attitudes of students to teachers, summarize teaching experience, improve subsequent teaching methods, and improve teaching quality. However, when evaluating teaching, random or malicious evaluations may occur among students, resulting in a large amount of noise in the evaluation data, which results in unsatisfactory feedback data. Therefore, this paper proposes a discrete Poisson mixture model to model the evaluation data of students with noise. Each discrete Poisson component in the mixture model corresponds to a class of students with similar evaluation modes. The model parameters in the loose distribution represent the evaluation scores in the corresponding evaluation mode. The log-likelihood function is constructed to measure the degree of fit between the mixed model and the evaluation data, and the gradient descent method is used to solve the model parameters with the highest degree of fit, to find the true evaluation of the students to the teacher, and to ensure the teacher-student relationship in the teaching evaluation system Communicate effectively. A large number of experimental results show that the model in this paper can quickly and accurately identify students with different evaluation modes from the evaluation data containing noise, and grasp the true evaluation of the students to teachers.

Key words: teaching evaluation system; crowdsourcing ideas; Poisson mixture model; parameter estimation method

0　引言

随着网络平台的普及，学校组织学生通过网络教学评价系统评价教师的课堂教学，已成为各大高校取代手工统计方式的普遍选择。学生评教作为高校教学质量评价的重要环节，越来越多的人开始探究如何借助这种方式进一步有效、科学地管理教学。教学评价主要是学生根据相应的评价指标对教师在这段时间内的教学状况进行打分，从而帮助教师总结教学经验、改进教学方法，达到最终促进教学中的师生沟通、提高教学质量的目的。

然而，目前大多数高校所采用的教学评价系统中，存在着参与性低及恶意评价的主要问题。学生缺乏参与评价的主动性，认为自己的评价不会改变课程的教授方式，即使许多学校强制进行教学评价，学生在进行评价时抱着完成任务的心态对所有老师随意打分或故意给老师打出低分的情况也常常能够看到。如此，影响了教师授课的积极性，进而造成了教学效果的停滞不前，同时学生认为教学评价没有实际作用，导致了恶性循环。

由于收集的学生评价数据通常是包含噪声的，即随机给出的、恶意评价的，没有直接表达他们的真实想法，难以通过简单的多数性投票策略来获取学生对于教师的真实评价，需要进一步地从这些包含噪声的数据中分析出学生的真实情况。如果将学生视为参与众包的众包工人，通过收集大量学生对教师的反馈情况实现学生的广泛参与的这种评价行为与众包服务相似，可以利用从众包任务获取真实标签的方法来处理学生的评价数据。但是进行众包数据处理时，需要非常复杂的参数模型来建模众包工人的贴标能力，采用类似 EM 的算法进行参数更新求解，容易陷入到局部最优，无法准确便捷地获取真实的任务标签。

为了避免上述缺陷，本文建议使用离散泊松混合模型对含有噪声的学生评价数据进行建模，将具有相似评价行为模型的学生对应于一个离散泊松模型，使用模型中的参数来表示学生的具体评价分数，再构建最大似然函数来评估模型和评价数据的拟合程度，通过梯度下降的方法找到使得拟合程

收稿日期：2022-01-26；**修回日期：**2022-04-01　**基金项目：**国家自然科学基金资助项目(61976163，61902284)

作者简介：黄浩(1986-)，男，湖北潜江人，研究员，博导，博士，主要研究方向为数据挖掘；颜钱(1994-)，男，湖北大悟人，博士研究生，主要研究方向为数据挖掘；甘庭(1989-)，男(通信作者)，湖北武汉人，讲师，博士，主要研究方向为约束求解(ganting@whu.edu.cn)；李石君(1964-)，男，湖南岳阳人，教授，博导，博士，主要研究方向为数据库、数据挖掘、大数据.

度最高时的模型参数，从而识别出具有不同评价模式的学生，确定学生对于教师的真实评价情况。大量实验表明，本文模型能够快速准确地从含有噪声的评价数据中识别出不同类别打分的学生，同时掌握学生对教师教学工作的准确反馈。

1　相关工作

教学评估主要是收集学生对于所上课程的评价来分析学生对于课程的真实态度，涉及的相关研究主要是对于评价数据的收集和处理。

当前的评价数据收集工作的主要途径是网络调查服务，但是学生的评价会由于受到调查疲劳或者学生对取得课程成绩的满意程度的影响，导致收集的数据是低回复或者低质量的。因此当前有些教学评估研究集中于获取优质的评价数据，通过利用人工智能技术实现虚拟的会话代理服务[1]与学生进行个人访谈，产生更加高质量的评价数据。还有一些研究人员通过额外的数据平台获取教师的相关数据[2,3]作为评价数据的补充来实现更加完整的教学评估。

对于开放性问题的文本评价数据的处理主要通过相应算法对学生回复文本进行主题检测，或将其分类为类别或情绪。许多研究人员利用 LDA 主题模型从学生书面反馈中提取主题[4,5]，与聚类模型相比，LDA 模型可以为评论找到更多相关主题，再基于相应的分类技术，将学生评论分类为正面或负面评论，从而更好地获取学生的情绪态度。也有研究人员直接利用自然语言处理的相关技术对于学生的评论文本进行分析[6,7]，捕获有意义的情绪信息，为了避免收集的评论回复中存在异常信息对分析结果准确性的影响，一些评估方法也使用了基于神经网络的异常检测算法[8]来提升算法的效果。

对于一些客观问题的评分数据的处理与众包任务相似，也是本文重点研究的内容，下面将回顾众包模式研究的相关工作，对众包的出现及现有研究的不足进行介绍。

众包并不是一个新现象，然而，近年来，互联网企业商业模式的巨大成功，又引起了大家对众包的关注。众包是一种通过互联网外包和利用分布式人工计算能力来解决特定功能集的方法[9]，人类主动或被动地参与计算过程，尤其是对于人类本质上比计算机更容易完成的任务[10]。众包主要有两大发展模式，即整合型众包和选择型众包[11]。整合型众包是指每一个单独的个体所带来信息的作用是微乎其微的，然而众多个体信息整合的结果可以带来巨大的价值。选择型众包即在众多解决方案中只存在一个最优的满足要求会被采纳，而其他的会被淘汰。

尽管众包的概念起源于商业，但是其应用已超越了商业，被广泛地应用于各个其他的领域。在计算机方面，许多学者使用众包来支持他们在数据采集、数据清洗、质量评估等方面的工作[12]，还有交通领域所提出的众包交通检测、众包配送等概念，还有许多其他如图书情报领域等方面。维基百科是众包应用中成功的案例，开创了一种人人参与知识创造和积累的运作模式。然而，众包模式中也存在着一些风险，其中包括较差的任务质量、不诚信的参与者以及众包过程中因参与者数量多而造成的不可控性等。因此，未来的研究中也会关注到众包的风险管理。本文中，也是就众包在教学评价中存在的不足所提出的解决方案。

在教学评价系统中，通过众包服务的思想收集学生对教师教学工作打分的标签结果，分析标签结果中反映的学生认同度来对教师的教学进行反馈。通过众包服务收集标签已被证明在许多应用中是有效的，例如自然语言处理和医疗数据处理。收集到的标签通常是有重复的，即不同的标记者为同一个实例提供了多个可能相互矛盾的标记结果。这种重复标签的方案在标签成本与数据质量之间取得了良好的平衡，并引起了对多个不可靠来源建模数据的研究课题的关注。

对于众包中的标签任务，最常见的目标是标签预测，即获取实例的可靠标签。为此，主流方法假设每个实例都存在真实结果标记，并尝试根据标记者给出的标签预测真实标签。Karger 等人[13]提出了一种算法，该算法在实例和标记者之间迭代地传递消息；Liu 等人[14]通过引入标记者的先验知识并使用图模型的变型方法来推断相应的生成模型的方式推广该算法。Whitehill 等人[15]提出了不同标记着具有不同的能力以及实例的争议性问题，这些都是通过概率模型与真实标签一起推断出来的。此外，一些新的技术，例如噪声校正[16]和不平衡学习[17]被用来提高标签的质量，尤其是在实例中带有少量噪声标签的情况下。

当实例可以在向量空间中被表示时，一个密切相关的主题是从标签中学习分类器。可以通过首先使用上述标签预测技术推断真实标签，然后通过传统分类方法学习分类器来轻松完成此任务。更复杂的方法包括直接从标记者给出的标签中学习，同时推断隐藏的标记者能力[18]，将标记者看做与最终分类器有关的个人分类器[19]，并将标记者的能力建模为实例空间的函数，并与最终分类器一起推断参数。这些工作以不同的方式对标记者的能力进行了建模，但是当他们将实例空间看做一个整体时，并没有明确地涉及实例的争议性问题。这种工作的一个缺点是，对于许多现实世界中的任务而言，实例的向量形式并不总是很容易获得的。

尽管大多数现有的工作旨在针对每个实例预测一个可靠的标签[20]，但仍有一些人尝试从其他方面解决问题。Wang 和 Zhou[21]提出了一个通用的理论框架来帮助从有高准确性的高质量标记者中识别(或消除)低质量的标签。Welinder 等人[22]提出了“思想学校”的概念，该概念允许标记者标记和提取不同的观点；Tian 和 Zhu[23]通过聚类标记者的标记结果估计了标记者的能力和实例的争议性，从而扩展了这个概念。Ertekin 等人[24]研究了仅通过查询一部分标记者来估计标记者的主要意见的近似人群问题。

由于众包带来的巨大机遇，许多的研究人员开发了大量技术来处理众包学习中的不精确性、随机性和不确定性问题[25]。但是，大多数现有工作都仅仅涵盖了某些方面。相反，本文的模型预测标记者本身的数据生成而不是某些方面，并且其功能足以包含标记者的行为模式和不同意见，同时保持模型本身的简单灵活。

2　相关概念

2.1　混合模型

混合分布模型的出现解决了用单一模型来研究问题的不足，它的本质就是融合几个单分布模型，来使得模型更加复杂，从而产生更复杂的样本，以此解决单一模型无法产生的样本的情况。假设随机变量 $x=(x_1,\dots,x_N)$ 来自由 M 个总体 $G_1,\dots,G_m$ 分别以比例 $\lambda_1,\dots,\lambda_m$ 混合而成的分布 G，于是 $f(x\mid\Lambda,\Theta)$ 的密度函数可以表示为

$$f(x\mid\Lambda,\Theta)=\lambda_1 f_1(x\mid\theta_1)+\cdots+\lambda_M f_M(x\mid\theta_M)$$

其中 $\sum_{i=1}^{M}\lambda_i=1,\lambda_i\ge 0,i=1,\dots,M$，$f(x\mid\theta_j)$ 和 θ_j 分别是相应于总体 G_j 的密度函数和参数，$\Lambda=(\lambda_1,\dots,\lambda_M)$，$\Theta=(\theta_1,\dots,\theta_M)$。称随机变量 x 服从混合模型 $f(x\mid\Lambda,\Theta)$。$f_m(x\mid\theta_m)(m=1,\dots,M)$ 是第 m 个分模型的概率密度函数，可以看成是选定第 m 个模型后，该模型产生 x 的概率；$\lambda_m(m=1,\dots,M)$ 是第 m 个分模型的权重，可看做第 m 个分模型的先验概率，调整权重 λ_m 将极大地影响混合模型的概率密度函数曲线，因此通过调整权重，混合模型便可以拟合更复杂更多变的样本。

混合模型是一个灵活且强有力的概率建模工具，在理论和实践中得到了极为广泛的应用，因为它具有以下优势：1)

混合模型提供了用简单的结构模拟复杂分布的一个有效的模型。比如，正态分布是现实生活中最常见、最重要的分布，因此，应用也最为广泛，许多随机现象当样本量足够大时都可以用正态分布逼近，理论可证明，利用混合正态分布(也称为混合高斯分布)可以逼近任何一个光滑分布，即只要项数 M 足够大，它们之间的权重设定地足够合理，混合分布模型可以用于描述复杂现象。因此，混合模型有助于解决实际生活中的许多复杂问题。2)混合模型所提供的模拟较为自然。当 M=1 时，模型为单一分布，则说明数据具有相同的性质；当 M>1 时，则说明数据为来自不同分布的混合数据，具有不同的性质，因此，在聚类分析、判别分析等领域中都有着广泛的应用。

2.2　隐变量

在混合分布模型中，存在着未知的数据，称为隐变量。

可以设想观测到的随机变量 $x=x_1,\ldots,x_N$ 是这样产生的：首先依概率 $\lambda_m(m=1,\ldots,M)$ 选择第 m 个分布 G_m，然后依这个成分的概率分布 $f_m(x|\theta_m)$ 生成观测数据 $x_n=(n=1,\ldots,N)$，N 个观测数据中可能有多个来自于同一个成分。这时，观测数据 $x=x_1,\ldots,x_N$ 是已知的，而反映观测数据 x_n 来自于总体分布的哪一个成分是未知的，即隐变量，用 γ_{nm} 表示有：

$$\gamma_{nm}=\begin{cases}1,第n个观测变量来自第m个成分\\0,否则\end{cases}$$

其中，$n=1,\ldots,N;m=1,\ldots,M$。

3　模型框架

在教学评价系统中，学生评教是最直接、真实和可靠的，因为学生是教师教学效果的直接体现者。学生评价教师的教学情况，其打分者是学生，打分对象是教师的教学工作，通过教学评价系统，共同促进教学工作的实施与改进。在接下来的介绍中，首先对问题进行描述，然后介绍用来概率拟合的离散泊松混合模型并对模型进行解释，最后提出参数估计方法以及根据参数进行教师教学质量分析。

3.1　问题描述

假设有 S 个学生为 N 个教师进行教学工作评分，每个学生的评分取值于集合{1,2,3,4,5,6,7,8,9,10}，分值越高代表对于教学工作的满意程度越高。所有学生的打分使用集合 $Y=\{\{y_{si}\}_{i=1}^N\}_{s=1}^S$ 表示，其中 $y_{si}\in\{1,2,3,4,5,6,7,8,9,10\}$ 表示第 $s\in\{1,\ldots,S\}$ 个学生对第 $i\in\{1,\ldots,N\}$ 个教师的评价分数。教学评估的问题是根据给定的打分集合 Y 找到不同类别打分模式的学生以及对于每个教师教学工作的真实评价。

为了生成一个简单但更灵活和更实用的模型，试图直接理解和模拟打分结果集合 Y 的生成过程。这是因为教师具有丰富的教学经验，对于所教课程会具有相对稳定的课程知识输出和课堂教学表现，所以学生的评价对象具有相对固定的教学模式，虽然不同类别的学生可能这种教学模式有不同的接受情况，但是同类别的学生可能会产生类似的教学评价，就可以利用生成式的机器学习模型来找到这种教学模式对不同类别学生产生的效果评价。为此，将一个学生视为一个单元，对于每一个学生 s 模拟他对所有任务打分结果标签的概率分布 $\{y_{s1},\ldots,y_{sN}\}$，将分布表示为 $p(\boldsymbol{y}|\boldsymbol{\theta})$ 并假设每个 $\boldsymbol{y}_s=[y_{s1},\ldots,y_{sN}]^T$ 来自这个分布。这样，需要做的就是为 $p(\boldsymbol{y}|\boldsymbol{\theta})$ 选择合适的模型形式，并根据 Y 中观察到的标签推导出相应的模型参数 θ。

3.2　离散泊松混合模型

本文选择 N 维离散泊松混合模型来拟合变量 $\boldsymbol{y}$ 的概率分布。其原因有两个：1)每个打分结果标签的值受离散泊松分布的影响。2)变量 $\boldsymbol{y}$ 是 N 维的，因为它反映了给定 N 个打分任务上每个学生的打分结果。在不失一般性的情况下，本文假设离散泊松混合模型可以描述为

$$p(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\alpha})=\sum_{k=1}^{K}\alpha_k\prod_{i=1}^{N}\frac{\mu_{ki}^{y_i}}{y_i!g(\mu_{ki})}$$

其中，$\boldsymbol{\mu}=\{\mu_1,\ldots,\mu_K\}$，$\mu_k=(\mu_{k1},\ldots,\mu_{kN})$ 是指第 k 个泊松成分的泊松参数，$\boldsymbol{\alpha}=\{\alpha_1,\ldots,\alpha_K\}$，$\alpha_k$ 是第 k 个泊松成分的系数(或权重)，$g(\mu_{ki})$ 是离散泊松分布的归一化因子，定义如下

$$g(\mu_{ki})=\sum_{v=1}^{10}\frac{\mu_{ki}^{v}}{v!}$$

在该模型中，每个学生 s 的打分结果被看做是由泊松成分生成的。如果一群学生的行为相似(例如，这些学生都不认真打分随机地给出打分结果，学生故意进行恶意评分扰乱结果等)，他们的打分结果往往由相同的泊松成分生成。此外，如果该群体的学生占有很大比例，则他们相应的泊松成分的系数 α_k 将大于其他的泊松成分系数。同时本文不假设影响因素如何共同造成打分集合 Y 的不确定性和错误，直接模拟 Y 的生成过程，并将影响因素带来的影响融入到泊松参数 μ_k 中。

3.3　模型参数估计

根据学生打分集合 $Y=\{\{y_{si}\}_{i=1}^N\}_{s=1}^S$ 构建关于上述离散泊松混合模型参数 $\boldsymbol{\mu}$ 和 $\boldsymbol{\alpha}$ 的似然函数为

$$p(Y|\boldsymbol{\mu},\boldsymbol{\alpha})=\prod_{s=1}^{S}\left(\sum_{k=1}^{K}\alpha_k\prod_{i=1}^{N}\frac{\mu_{ki}^{y_{si}}}{y_{si}!g(\mu_{ki})}\right)$$

其对数似然函数为

$$\ln(p(Y|\boldsymbol{\mu},\boldsymbol{\alpha}))=\sum_{s=1}^{S}\ln\left(\sum_{k=1}^{K}\alpha_k\prod_{i=1}^{N}\frac{\mu_{ki}^{y_{si}}}{y_{si}!g(\mu_{ki})}\right)$$

需要最大化上述对数似然函数，得到最优的离散泊松混合模型参数 $\boldsymbol{\mu}^*$ 和 $\boldsymbol{\alpha}^*$，即

$$\boldsymbol{\mu}^*,\boldsymbol{\alpha}^*=argmax_{\boldsymbol{\mu},\boldsymbol{\alpha}}\ln(p(Y|\boldsymbol{\mu},\boldsymbol{\alpha}))$$

将上述的对数似然函数分别对 $\boldsymbol{\mu}$ 和 $\boldsymbol{\alpha}$ 的每个分量求偏导，有

$$\frac{\partial\ln(p(Y|\boldsymbol{\mu},\boldsymbol{\alpha}))}{\alpha_k}=\sum_{s=1}^{S}\frac{\prod_{i=1}^{N}\frac{\mu_{ki}^{y_{si}}}{y_{si}!g(\mu_{ki})}}{\sum_{m=1}^{K}\left(\alpha_k\prod_{i=1}^{N}\frac{\mu_{mi}^{y_{si}}}{y_{si}!g(\mu_{mi})}\right)}$$

$$\frac{\partial\ln(p(Y|\boldsymbol{\mu},\boldsymbol{\alpha}))}{\mu_{ki}}=\sum_{s=1}^{S}\frac{\alpha_k\prod_{j\neq i}\frac{\mu_{kj}^{y_{sj}}}{y_{sj}!g(\mu_{kj})}\times\frac{y_{si}\mu_{ki}^{y_{si}-1}g(\mu_{ki})-\mu_{ki}^{y_{si}}g'(\mu_{ki})}{y_{si}!g^2(\mu_{ki})}}{\sum_{m=1}^{K}\left(\alpha_m\prod_{j=1}^{N}\frac{\mu_{mj}^{y_{sj}}}{y_{sj}!g(\mu_{mj})}\right)}$$

其中 $g'(\mu_{kj})$ 是函数 $g(\mu_{kj})$ 关于其自变量的导函数，定义如下

$$g'(\mu_{kj})=\sum_{v=1}^{9}\frac{\mu_{ki}^{v}}{v!}+1$$

利用梯度下降法，选取 $\boldsymbol{\mu}$ 和 $\boldsymbol{\alpha}$ 的初值 $\boldsymbol{\mu}^0$ 和 $\boldsymbol{\alpha}^0$，再利用上述的偏导，按照如下方法更新 $\boldsymbol{\mu}^t$ 和 $\boldsymbol{\alpha}^t$ 的值得到 $\boldsymbol{\mu}^{t+1}$ 和 $\boldsymbol{\alpha}^{t+1}$，重复迭代直到收敛，

$$\mu_{ki}^{t+1}=\mu_{ki}^{t}+\theta^t\frac{\partial\ln(p(Y|\boldsymbol{\mu}^t,\boldsymbol{\alpha}^t))}{\mu_{ki}}$$

$$\alpha_{k}^{t+1}=\alpha_{k}^{t}+\theta^t\frac{\partial\ln(p(Y|\boldsymbol{\mu}^t,\boldsymbol{\alpha}^t))}{\alpha_{k}}$$

其中 θ^t 是第 t 步的更新步长，为了使迭代终止，本文要求 θ^t 满足如下要求，

$$\lim_{t\to+\infty}\theta^t=0,\sum_{t=1}^{\infty}\theta^t=\infty$$

本文中取 $\theta^t=\frac{1}{t}$。

3.4　泊松成分分析

在 3.3 节中，本文利用观测到的部分学生打分情况，对 3.2 节中提出的离散泊松混合模型的参数进行估计。对于任意一位学生 s，其打分为 y_s，该学生的打分与离散泊松混合模型中第 k 个成分的关联度可以由下面似然函数估计：

$$r(y_s,k)=\frac{\alpha_k\prod_{i=1}^{N}\frac{\mu_{ki}^{\ y_{si}}}{y_{si}!g(\mu_{ki})}}{\sum_{m=1}^{K}\alpha_m\prod_{i=1}^{N}\frac{\mu_{mi}^{\ y_{si}}}{y_{si}!g(\mu_{mi})}}$$

3.5　教师教学质量分析

得到教学评分的离散泊松混合模型参数估计后，可根据模型对各教学任务进行教学质量分析。现考虑第 i 个教学任务的得分 y_i，可估计 $p(y_i=m\mid\boldsymbol{\mu},\boldsymbol{\alpha})$ 如下

$$\begin{aligned}
&p(y_i=m\mid\boldsymbol{\mu},\boldsymbol{\alpha})\\
&=\sum_{\overline{y_l}}p\left(y_i=m,\overline{y_l}\mid\boldsymbol{\mu},\boldsymbol{\alpha}\right)\\
&=\sum_{\overline{y_l}}\sum_{k=1}^{K}\alpha_k\prod_{j=1}^{N}\frac{\mu_{kj}^{\ y_j}}{y_j!g(\mu_{kj})}=\sum_{k=1}^{K}\alpha_k\sum_{\overline{y_l}}\prod_{j=1}^{N}\frac{\mu_{kj}^{y_j}}{y_j!g(\mu_{kj})}\\
&=\sum_{k=1}^{K}\alpha_k\underbrace{\sum_{y_1=1}^{10}\cdots\sum_{y_N=1}^{10}\prod_{j=1}^{N}\frac{\mu_{kj}^{\ y_j}}{y_j!g(\mu_{kj})}}_{\text{不包含}y_i\text{剩下的}N-1\text{个}y_j\text{求和}}\\
&=\sum_{k=1}^{K}\alpha_k\frac{\mu_{ki}^{\ y_i}}{y_i!g(\mu_{ki})}\underbrace{\sum_{y_1=1}^{10}\frac{\mu_{k1}^{\ y_1}}{y_1!g(\mu_{k1})}\cdots\sum_{y_N=1}^{10}\frac{\mu_{kN}^{\ y_j}}{y_j!g(\mu_{kN})}}_{\text{不包含}y_i\text{剩下的}N-1\text{个}y_j\text{求和}}\\
&=\sum_{k=1}^{K}\alpha_k\frac{\mu_{ki}^{\ m}}{m!g(\mu_{ki})}
\end{aligned}$$

其中 $\overline{y_l}$ 是 $(y_1,\dots,y_N)$ 去掉变量 y_i 之后的 N-1 个教学任务的得分，$m\in\{1,\dots,10\}$。可以得到第 i 个教学任务的得分估计如下按照概率最大对应的得分 m 作为该教学任务的得分或者取期望

$$y_i^*=arg\max_m p(y_i=m\mid\boldsymbol{\mu},\boldsymbol{\alpha})$$

4　实验结果与分析

本节首先介绍实验用到的数据集和评价指标，然后在数据集上验证本文的模型在以下两个方面是有效的，即 1)它可以按学生的行为模式对他们进行分类，2)它可以准确地预测真实标签。

4.1　实验设置

在进行教学评估时，每个学生收到的教学评估调查问卷包含的主要内容如下表 1 所示，包含所上课程的时间、课头号、课程名称、课程号、授课教师、评教分数等内容，学生根据在上课时的课程体验对该课程进行评分，评分取值可从 1~10 分进行选择。本文实验使用的评分数据是由 100 名学生对 50 名教师模拟打分产生，首先给每个教师设定一个真实的评分标签，然后将学生分为三类进行打分：正常打分的学生是根据给定的教师分数标签进行上下浮动打分，随机打分的学生在 0~10 之间随机打分，恶意打分的学生在 0~5 分之间恶意打分。

实验评价包含两个指标:

1)根据模拟生成数据时的学生类别计算每个学生对于占比最大的泊松分布分量的隶属度，统计每一个类别学生的隶属度均值，以正常学生类别为例计算隶属度均值的方式为

$$membership_{normal}=\frac{\sum_{s=1}^{S}r(y_s,k)\times I(s\ is\ normal)}{\sum_{s=1}^{S}I(s\ is\ normal)}$$

其中 $r(y_s,k)$ 表示学生 s 对于最大泊松分布分量 k 的隶属度；$I(.)$是一个指示函数，当括号内的条件成立时取值为 1，否则取值为 0。

2)根据计算得到的泊松分布参数来获取教师的分数标签，与给定的真实标签的差异。假设评分数据中包含 N 名教师，每个教师 i 的真实标签为 y_i，对应预测的教师分数标签为 p_i，计算标签的差异得分为

$$score=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{(abs(y_i-p_i)+1)!}$$

表 1　教学评估问卷表

Tab. 1　Teaching assessment questionnaire

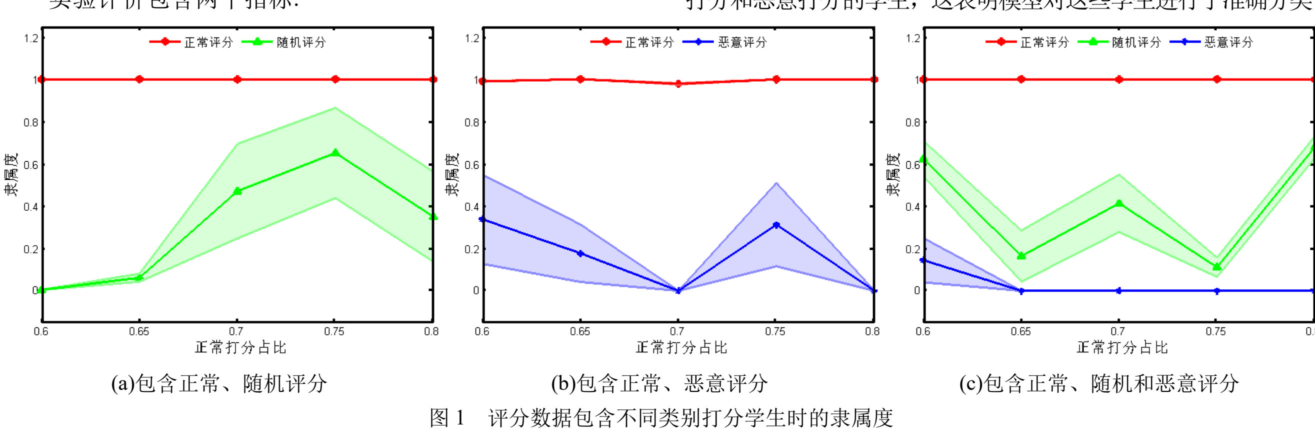
课程时间	2020-2021 学年第一学期
课头号	20201021082
课程名称	科技写作
课程号	3350520011037
授课教师	张三
评教分数	9

4.2　结果评估

4.2.1 学生分类

为了验证本文模型对具有不同行为模式的学生进行分类的能力，本文首先给与每个教师一个随机生成的分数标签，然后根据该标签生成模拟打分数据，在测试评评分数据中除了正常评分，还包含随机评分、恶意评分或者随机和恶意评分都存在这三种情况下，不同类别打分学生在主要离散泊松成分中的隶属度的平均期望。假设评分数据中正常评分学生所占比例为 α(α 取值从 0.6 到 0.8 之间变化)，当剩余评分学生只有随机评分或者恶意评分时，对应随机或者恶意评分学生所占比例为 1-α，当剩余评分学生包含随机评分和恶意评分时，对应随机评分和恶意评分所占比例均为(1-α)/2。为确保实验的可信度，模拟生成多组实验数据，记录每次执行的平均期望，并报告多次运行结果的均值和方差(通过结果图中的阴影区域标识)。

图 1 展示了当评分数据包含不同类别打分学生时，各类学生在主要离散泊松成分中的隶属度结果。从中可以观察到，主要离散泊松成分中进行正常打分的学生的平均隶属度显著高于随机打分和恶意打分的学生，这表明模型对这些学生进行了准确分类。



(a)包含正常、随机评分　(b)包含正常、恶意评分　(c)包含正常、随机和恶意评分

图 1　评分数据包含不同类别打分学生时的隶属度

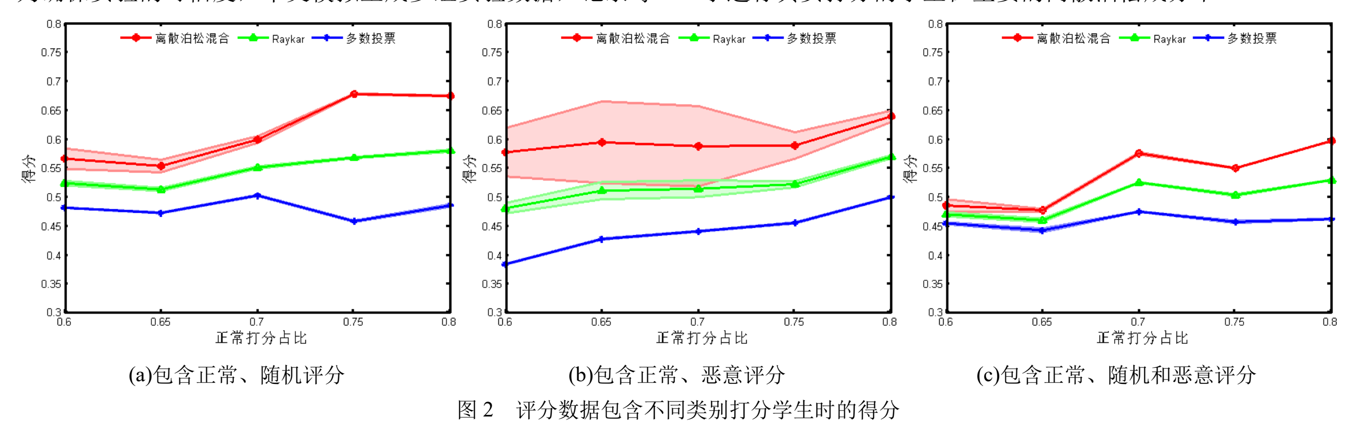
Fig. 1　The grading data contains the degree of membership when scoring students in different categories

4.2.2 标签预测

为了验证本文模型能够从评分数据中心分析出学生对于教师的真实评价，首先给与每个教师一个随机生成的分数标签，然后根据该标签生成模拟打分数据，在测试评分数据中除了正常评分，还包含随机评分、恶意评分或者随机和恶意评分都存在这三种情况下，本文模型找到的教师评价分数和

真实分数标签的差异，同时也报告了采用多数投票策略和 Raykar 等人提出的标签预测算法[12]得到的教师评价分数和真实分数标签之间的差异作为对比。假设评分数据中正常评分学生所占比例为 α(α 取值从 0.6 到 0.8 之间变化)，当剩余评分学生只有随机评分或者恶意评分时，对应随机或者恶意评分学生所占比例为 1-α，当剩余评分学生包含随机评分和恶意评分时，对应随机评分和恶意评分所占比例均为(1-α)/2。为确保实验的可信度，本文模拟生成多组实验数据，记录每次执行的平均期望，并报告多次运行结果的均值和方差(通过结果图中的阴影区域标识)。

图 2 展示了当评分数据包含不同类别打分学生时，找到的教师评价分数和真实分数标签之间的差异得分，可以看出本文模型找到的教师评价分数与两个对比策略相比具有更高的准确度，背后的原因是在本文的模型中，随机打分的学生或恶意打分的学生被分配到了非主要的离散泊松成分，留下了进行真实打分的学生在主要的离散泊松成分中。



(a)包含正常、随机评分 (b)包含正常、恶意评分 (c)包含正常、随机和恶意评分

图 2 评分数据包含不同类别打分学生时的得分

Fig. 2 The scoring data contains the scores of students in different categories

5 结束语

本文提出了一种离散泊松混合模型来模拟学生对于教师教学工作的打分结果，并提出了一种梯度下降的方法用于模型的参数估计。该模型直接模拟打分结果的生成过程，无须在学生打分能力和实例争议性等影响因素上进行额外的假设或推断。在实验结果上证明了本文模型在标签预测和对不同行为模式的学生进行分类方面的有效性。同以前的教学评价结果评估相比，本文模型具有更高的容错性，即使存在随机打分、恶意打分学生，也能够得到可靠的评估结果，对教师的教学工作有一个准确的反馈，反映教学中的真实情况。

参考文献：

[1] Wambsganss T, Winkler R, Söllner M, *et al.* A conversational agent to improve response quality in course evaluations [C]// Proc of ACM CHI. New York: ACM Press, 2020: 1-9.

[2] Kavalchuk A, Goldenberg A, Hussain I. An empirical study of teaching qualities of popular computer science and software engineering instructors using ratemyprofessor. com data [C]// Proc of the 42th International Conference on Software Engineering: Software Engineering Education and Training. Piscataway, NJ: IEEE Press, 2020: 61-70.

[3] Lin Q, Zhu Y, Lu H, *et al.* Improving university faculty evaluations via multi-view knowledge graph [J]. Future Generation Computer Systems, 2021, 117: 181-192.

[4] Gottipati S, Shankararaman V, Lin J. Latent Dirichlet Allocation for textual student feedback analysis [C]// Proc of the 26th International Conference on Computers in Education. 2018: 220-227.

[5] Unankard S, Nadee W. Topic detection for online course feedback using LDA [C]// Proc of the 4th International Symposium on Emerging Technologies for Education. Berlin: Springer, 2019: 133-142.

[6] Andersson E, Dryden C, Variawa C. Methods of applying machine learning to student feedback through clustering and sentiment analysis [C]/ Proc of CEEA. 2018. https://doi. org/10. 24908/pceea. v0i0. 13059

[7] Onan A. Mining opinions from instructor evaluation reviews: a deep learning approach [J]. Computer Applications in Engineering Education, 2020, 28 (1): 117-138.

[8] Mao Y, Zhu Y, Zhang S, *et al.* Detecting interest-factor influenced abnormal evaluation of teaching via multimodal embedding and priori knowledge based neural network [C]// Proc of IEEE ISPA/BDCloud/SocialCom/SustainCom. Piscataway, NJ: IEEE Press, 2019: 1201-1209.

[9] Bhatti S S, Gao X, Chen G. General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey [J]. Journal of Systems and Software, 2020, 167: 110611.

[10] Tong Y, Zhou Z, Zeng Y, *et al.* Spatial crowdsourcing: a survey [J]. The VLDB Journal, 2020, 29 (1): 217-250.

[11] 霍绪艳. 中国众包发展现状研究 [J]. 商情, 2011 (47): 117-117. (Huo Xuyan. Research on the status quo of crowdsourcing development in China [J]. ShangQing, 2011 (47): 117-117.)

[12] Raykar V C, Yu S, Zhao L H, *et al.* Learning From Crowds [J]. Journal of Machine Learning Research, 2010, 11 (2): 1297-1322.

[13] Karger D R, Oh S, Shah D. Iterative learning for reliable crowdsourcing systems [C]// Proc of the 24th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2011: 1953-1961.

[14] Liu Q, Peng J, Ihler A. Variational inference for crowdsourcing [C]// Proc of the 25th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 692-700.

[15] Whitehill J, Ruvolo P, Wu T, *et al.* Whose vote should count more: optimal integration of labels from labelers of unknown expertise [C]// Proc of the 22th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2009: 2035-2043.

[16] Zhang J, Sheng V S, Wu J, *et al.* Improving label quality in crowdsourcing using noise correction [C]// Proc of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1931-1934.

[17] Zhang J, Wu X, Sheng V S. Imbalanced Multiple Noisy Labeling [J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27 (2): 489-503.

[18] Yan Y, Rosales R, Fung G, *et al.* Modeling Multiple Annotator Expertise in the Semi-Supervised Learning Scenario [J]. Computer Science, 2012: 674-682.

[19] Kajino H, Tsuboi Y, Kashima H. Clustering crowds [C]// Proc of the 27th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2013,

27 (1): 1120-1127.

[20] Zhao Y, Zhu Q. Evaluation on crowdsourcing research: Current status and future direction [J]. Information Systems Frontiers, 2014, 16 (3): 417-434.

[21] Wang W, Zhou Z H. Crowdsourcing label quality: a theoretical analysis [J]. Science China Information Sciences, 2015, 58 (11): 1-12.

[22] Welinder P, Branson S, Belongie S, *et al.* The multidimensional wisdom of crowds [C]// Proc of the 23th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2010: 2424-2432.

[23] Tian Y, Zhu J. Learning from crowds in the presence of schools of thought [C]// Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 226-234.

[24] Ertekin S, Rudin C, Hirsh H. Approximating the crowd [J]. Data Mining and Knowledge Discovery, 2014, 28 (5-6): 1189-1221.

[25] Sheng V S, Zhang J. Machine learning with crowdsourcing: A brief summary of the past research and future directions [C]// Proc of the 33th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2019, 33 (01): 9837-9843.